ACOUSTIC VARIABILITY IN SPONTANEOUS CONVERSATIONAL SPEECH OF AMERICAN ENGLISH TALKERS

Ann K. Syrdal

AT&T Network and Computing Services, Holmdel, NJ

ABSTRACT

Speaker variability strongly impacts human perception and technology performance, yet large-scale, systematic study of the acoustic characteristics involved is rarely undertaken. This study provides statistics on selected segmental and suprasegmental acoustic parameters from measures made on spontaneous conversational telephone speech from 160 speakers in the Switchboard Corpus.[1] Since spontaneous conversational speech is more dynamically variable than read speech and is representative of actual human communication, it was preferred for our applied research purposes.

1. METHOD

1.1. Speech Samples

Ten male and ten female speakers were selected from each of eight American English dialect categories based on geographical location(s) during their first ten years of life. Whether or to what extent a talker's speech was characteristic of a regional dialect was not verified, since acoustic variability representative of the population of American English speakers, not dialect *per se*, was the focus of the study. Male and female talkers and the eight dialect groups sampled were roughly equivalent in age and education. One of each speaker's 5-10 minute conversations was selected for analysis based on ratings of naturalness, echo, static, and background noise.

1.2. Acoustic Measurements

Segmental acoustic measurements included (1) formant and fundamental frequencies of vowels /i,a,u/ sampled from stressed syllables in words whose standard American English transcriptions (dictionary pronunciations) specify these vowels, and (2) voice onset time (VOT) measurements of consonants /p/ and /b/ sampled in word-initial position immediately preceding a stressed vowel.

Suprasegmental acoustic characteristics were measured from a 5-7 second excerpt from the conversation. The excerpt was a representative utterance of relatively continuous speech consisting of one or more complete phrases or sentences. Prosodic measures included intonational variables, speaking rate, speech activity, speech level, glottal characteristics, and long-term spectral shape.

Acoustic measurements were analyzed statistically by a series of repeated measures analyses of variance (ANOVAs) for effects of sex or dialect, or as an interaction between sex and dialect. Tukey Studentized Range Tests were performed to test specific contrasts. Descriptive statistics were computed, and correlations between various acoustic measures were studied.

2. RESULTS

2.1. Segmental Measures: Vowels

As expected, the fundamental frequency (F0) and first three formant frequencies (F1, F2, and F3) were significantly higher for females than for males in vowels /i/, /a/, and /u/. For /i/ and /a/ there was no significant effect of dialect on any of the four measures,. However, for /u/ there were significant dialect effects on F2 and F3, shown in Figure 1. A high F2 group was composed of South Midland, Southern, Mixed, North Midland, and Western dialects, and a low F2 group was composed of Northern and New England dialects. A significant sex by dialect interaction reflected the finding that NYC females resembled the low F2 group, while NYC males sampled were within the high F2 group. Southern talkers had a higher F3, and Northern talkers, a lower F3. The remaining dialects were intermediate and not significantly different in F3 from either of the two extremes. The dialect differences on F2 and F3 in /u/ reflect the observation that /u/ in Southern and related dialects of American English tends towards a rounded front vowel, whereas the Northern dialects retain /u/ as a rounded back vowel. This is an example of variability that can be very problematic for automatic speech recognition systems.

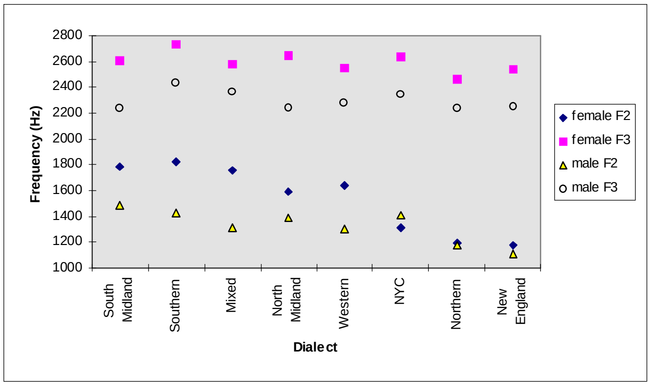

Figure 1. F2 and F3 of Vowel /u/ by Dialect and Sex

In a quantitative model of human vowel recognition,[2][3] a transformation of fundamental and formant frequencies from an acoustic to an auditory (critical band (Bark)) scale, and the representation of vowels as patterns of auditory distances between these components has been shown to normalize acoustic differences between men, women, and children talkers while preserving phonetically relevant information. A critical auditory distance of within 3 - 3.5 Barks was observed for a phonetically relevant auditory averaging effect,[4][5] and the same auditory distance was found to delimit some major phonetic dimensions of American English vowels.[2][3] A distance of less than 3 - 3.5 Bark between F1 and F0 was characteristic only of high vowels, and this critical distance between F3 and F2 was characteristic only of front vowels. Although not considered a major phonetic dimension of vowels, the 3 - 3.5 Bark distance between F2 and F1 marked only low vowels with compact spectra, such as /a/. These results were observed for vowels in read lists of isolated words, however, for which articulation is typically more precise and less variable than in spontaneous conversational speech. Auditory transformations of F0 - F3 vowel measures from the Switchboard Corpus were used to compare vowels in isolated word *versus* conversational contexts, and to compare some phonetically relevant vowel characteristics among normalized vowels spoken by the various talker groups.

Conversational vowels showed a tendency towards vowel reduction (that is, a divergence of formant patterns from those characteristic of a specific vowel category towards those characteristic of a less differentiated mid-central (schwa) vowel) when compared to vowels from isolated words. This was expected, since conversational speech is spoken more rapidly, is generally less carefully articulated, more intonationally varied, and more influenced by coarticulation with surrounding sounds than words spoken in isolated. A higher proportion of productions of the three vowels in conversational context crossed the critical distance boundaries of 3 - 3.5 Barks than did vowels spoken in citation form. However, the means across dialects for both female and male talkers were still on the expected side of the critical distance in all three auditory distance dimensions for vowels /i/ /a/ and /u/ in conversational contexts.

For the vowel /u/, there were significant effects of dialect on the auditory distances between F2 and F1 and between F3 and F2. F2-F1 distance, which was consistently greater than 3.5 Barks, was significantly larger for Southern and South Midland talkers than for New England and Northern talkers. F3-F2 distance was significantly larger for New England talkers (for which, typical of a back vowel, it consistently exceeded 3.5 Barks) than for South Midland talkers (for which, typical of a front vowel, it was within 3.5 Barks). With respect to F3-F2 distance for /u/, Northern and NYC talkers did not differ significantly from New Englanders, and Southern, Mixed, North Midland, and Western talkers did not differ significantly from South Midlands talkers. These results corroborate the previously noted tendencies towards a rounded front /u/ in Southern related dialects, and a rounded back /u/ in Northern related dialects.

Significant sex differences were observed for all three vowels normalized by auditory distance. F1-F0 distance was greater for females than males in the low vowel /a/, but was smaller for females than for males in high vowels /i/ and /u/. F2-F1 distance was smaller for females than for males in /a/, but greater for females than for males in /i/ and /u/. F3-F2 distance was smaller for female than male talkers in front vowel /i/, but did not differ significantly in back vowels /a/ and /u/. These results are consistent in that when the target auditory distance was within 3 - 3.5 Barks, female talkers always achieved a significantly smaller distance between spectral components than did males, and when the target auditory distance exceeded the critical distance, females never exhibited a significantly smaller distance than male talkers. These tendencies resulted in vowel categories that were more differentiated from one another for female talkers than for male talkers. Observations that female speech is typically more precisely articulated than male speech were also made in a study of Dutch vowel formant data from several speaking conditions[6] and in a study of American English using sentences from the TIMIT database.[7]

2.2. Segmental Measures: Consonants

There were no significant effects of sex or dialect or their interaction on Voice Onset Time (VOT) measures of /b/ and /p/. However, a significant dialect effect was found in a derived VOT range measure (VOT/p/ - VOT/b/, the difference in VOT between /p/ and /b/ measures). Western talkers made the largest differentiation between /p/ and /b/ in VOT, and their VOT range was significantly larger than NYC talkers, those with the smallest VOT range. Descriptive VOT statistics are listed in Table 1.

		VOT (ms)					
Talker Group	**N**	/b/	S.D.	/p/	S.D.	/p/ - /b/	S.D.
Females	80	-8	40	54	24	62	45
Males	80	-5	32	49	22	54	39
Western	20	-15	53	63	23	78*	63
Mixed	20	-21	46	53	36	74	55
North Midland	20	-2	19	58	20	61	25
Southern	20	-1	38	47	18	58	37
South Midland	20	-22	46	46	17	68	50
New England	20	7	17	53	24	45	30
Northern	20	2	20	46	17	44	23
NYC	20	9	7	46	18	38*	18

Table 1. Voice Onset Time (VOT) Measures (* denotes statistically significant ($p < 0.02$) dialect difference)

2.3. Prosody: Intonation

Measures of intonation included maximum F0, minimum F0, an estimate of reference line F0, and mean F0. Maximum, minimum, and reference line F0 (the frequency towards which a talker's F0 tends in the absence of tonal accents) were determined interactively from a display of the F0 contour. When automatically determined F0 estimates were questionable, they were verified by visual inspection of the speech waveform. Mean F0 was calculated from F0 estimates for each 5 ms frame that fell between the previously identified maximum and minimum F0 and for which the probability of voicing was at least 90%. Two additional derived F0 range variables were calculated: Maximum minus Reference, and Maximum minus Minimum.

Intonation measures for males and females are shown in Figure 2. All F0 measures and ranges were significantly higher for female talkers than for males. There were no significant dialect effects or interactions. Female talkers not only have higher-pitched voices than males for all intonation variables measured, but females have a much wider fundamental frequency range in conversational speech than male talkers do. For both sexes, the Maximum-Reference F0 range accounts for the majority (roughly three-quarters) of the entire (Maximum-Minimum) frequency range. Reference line F0 was generally lower in frequency than mean F0.

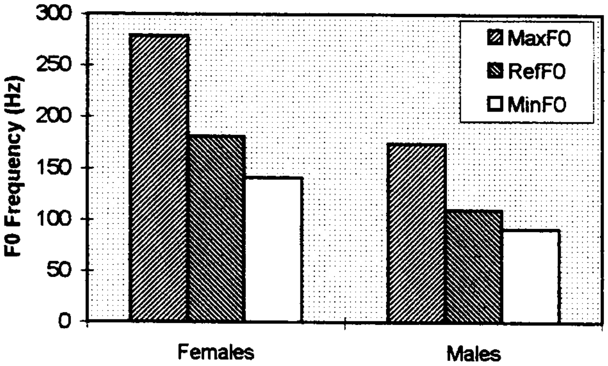

Figure 2. Intonation Measures of F0

2.4. Prosody: Speaking Rate

Descriptive statistics on speaking rate (words per second and syllables per second) are listed in Table 2. There were no significant effects of sex or dialect or their interaction on speaking rate.

Talker Group	Words/s.	S.D.	Syllables/s	S.D.
Females	3.68	0.61	4.77	0.72
Males	3.73	0.68	4.76	0.71
ALL	3.71	0.64	4.76	0.71

Table 2. Speaking Rate Measures

2.5. Prosody: Level and Dynamic Range

Mean active speech power was -24.4 dBm, with a standard deviation of 4.7 dB. The mean dynamic range measured from voiced segments within the 5-7 second excerpts was 13.9 dB, with a standard deviation of 4.0 dB. There were no significant sex or dialect effects or interactions for either measure.

2.6. Prosody: Glottal Characteristics

Glottal characteristics refer to acoustic properties attributable to variations in the manner in which the vocal folds of the glottis open and close during voicing. Two acoustic measures reflecting glottal characteristics were made on the conversational speech excerpts. These are H2-H1 (the level in dB of the second harmonic minus that of the first harmonic) and a count of the number of episodes of vocal creak.

H2-H1 has been used as an index of glottal configuration,[8] and for wide-band speech it is positive for laryngealized or "pressed" phonation, negative for breathy phonation, and approximately null for modal phonation. To minimize the influence of F1 on H1 and H2 amplitudes, research on glottal characteristics has focused on a highly restricted inventory, typically low vowels such as /a/, for which first formant frequency is maximal. However, a pilot analysis of 2 minute samples of wide-band spontaneous speech yielded H2-H1 measures very similar in the aggregate to those reported from restricted contexts. For the Switchboard excerpts, H2-H1 was measured interactively from a long-term FFT computed over the entire excerpt. H2-H1 was significantly higher for male talkers (mean = 13.7 dB, standard deviation = 7.7 dB) than for females (mean = 4.7 dB, standard deviation = 5.7 dB), and there were no dialect effects or interactions. In wide-band speech, H2-H1 is about 3 dB higher, on average, for male talkers than for females, but for telephone speech, low frequency attenuation introduced by the handset disproportionately raised H2-H1 for males, thus greatly increasing the sex difference. There was a highly significant negative correlation of -0.56 between mean F0 and H2-H1, but the effect was due primarily to a significant negative correlation (-0.31) for male talkers and to the group differences between male and female talkers. No significant correlation between mean F0 and H2-H1 was found for female talkers.

Vocal creak is a mode of phonation in which the vocal folds are held tightly together at one end so that only the other end is free to vibrate. This results in extremely laryngealized speech that is very low-pitched and of irregular periodicity and amplitude. Episodes of vocal creak were identified in spectrogram and waveform displays, and were verified auditorally. A count was made of the number of words in each excerpt spoken (at least in part) with creaky phonation. Significantly more instances of vocal creak were observed for female talkers (mean = 1.8) than for males (mean = 1.2). These results were surprising, since over all, males tend more towards laryngealized phonation than females.

2.7. Prosody: Spectral Shape

Long-term power spectra were determined for seven frequency bands, each 2.5 Barks in width, and then were normalized with respect to the least variable band in order to characterize spectral shape rather than absolute power. There was a significant sex by frequency band interaction because normalized power only differed between males and females in the lower two frequency bands, for which it was significantly lower for females.

3. SUMMARY

In general, significant differences were found between male and female speakers in measures related to frequency but not in those related to timing. Female vowels were more distinct from one another than were male vowels, but there were no sex differences in speaking rate. Female speakers used a much wider fundamental frequency range than males. There were significant sex differences in glottal characteristics and in long-term spectral shape, but no differences in overall level or dynamic range. Major dialect differences were observed only in the vowel /u/, for which there were basically Northern and Southern variants. A highly significant negative correlation was observed between mean F0 and overall H2-H1 (the level of the first harmonic subtracted from the level of the second harmonic). H2-H1 is an index of glottal configuration in wide-band speech, but in telephone speech it is also greatly influenced by low-frequency attenuation introduced by telephone sets.

4. REFERENCES

1. Godfrey, J. J., Holliman, E. C., & McDaniel, J. (1992). “SWITCHBOARD: Telephone speech corpus for research and development,” *IEEE Trans. ASSP I*, 517-520, San Francisco, March 1992.

2. Syrdal, A. K. (1985). “Aspects of a model of the auditory representation of American English vowels,” *Speech Communication*, **4**, 121-135.

3. Syrdal, A. K., and Gopal, H. S. (1986). “A perceptual model of vowel recognition based on the auditory representation of American English vowels,” *J. Acoust. Soc. Amer.*, **19**, 1086-1100.

4. Chistovich, L. A., and Lublinskaya, V. V. (1979). “The ‘center of gravity’ effect in vowel spectra and critical distance between the formants: Psychoacoustical study of the perception of vowel-like stimuli,” *Hear. Res.*, 185-195.

5. Traunmüller, H. (1981). “Perceptual dimension of openness in vowels,” *J. Acoust. Soc. Amer*. **69**, 1465-1475.

6. Koopmans-van Beinum, F. J. (1980). “Vowel contrast reduction: An acoustic and perceptual study of Dutch vowels in various speech conditions,” Ph.D. dissertation, Academic, Amsterdam.

7. Byrd, D. (1992). “Sex, dialects, and reduction,” In J. J. Ohala, T. M. Nearey, F. L. Derwing, M. M. Hodge, and G. E. Wiebe (Eds.), *Proceedings 1992 ICSLP*. Alberta, Canada: University of Alberta,. 827-830.

8. Klatt, D. H., and Klatt, L. C. (1990). “Analysis, synthesis, and perception of voice quality variations among female and male talkers,” *J. Acoust. Soc. Amer.*, **87**, 820-857.